Evolution Analysis of Call Ego-Networks

Shazia Tabassum and João Gama

LIAAD, Inesctec
University of Porto
Porto, Portugal
{up201402360@fe.up.pt,jgama@fep.up.pt}

Abstract. With the realization of networks in many of the real world domains, research work in network science has gained much attention now-a-days. The real world interaction networks are exploited to gain insights into real world connections. One of the notion is to analyze how these networks grow and evolve. Most of the works rely upon the socio centric networks. The socio centric network comprises of several ego networks. How these ego networks evolve greatly influences the structure of network. In this work, we have analyzed the evolution of ego networks from a massive call network stream by using an extensive list of graph metrics. By doing this, we studied the evolution of structural properties of graph and related them with the real world user behaviors. We also proved the densification power law over the temporal call ego networks. Many of the evolving networks obey the densification power law and the number of edges increase as a function of time. Therefore, we discuss a sequential sampling method with forgetting factor to sample the evolving ego network stream. This method captures the most active and recent nodes from the network while preserving the tie strengths between them and maintaining the density of graph and decreasing redundancy.

Keywords: Evolving Networks, Evolution Analysis, Call Networks, Densification, Ego-Networks, Forgetting Factor

1 Introduction

Enormous streams of graphs are generated by some of the real-time applications, at a speed of millions of nodes and billions of edges per day. Such social streams provide an abstraction of interactions between real world social entities or individuals. Studying the structural properties of these streams enables powerful insights and extrapolations of real world. Space and time complexity is one of the challenging issues related to analyzing these streams. Networks representing real world social structures are usually temporal and evolving. The rapidly changing and evolving structure of these graphs, calls for an exigency of latest and up to date results. Processing the real-time network stream as it arrives, is one of the best solutions for the above problem. Therefore, we employ the stream processing approach to process enormous data. Some of the social network analysis

methods that can be applied over streams of graphs are given in [12]. Furthermore, we use a streaming ego network approach over a telecommunications' call graph stream of temporal edge/calls' as in [14].

An ego network is based on the relationships of a single node called "ego" with the other nodes in a social network. An ego can represent an individual, entity, object or organization. All the other nodes related to ego in the network are called alters. An ego network maps the relationships of an ego with alters and also between themselves. In the recent work [4] by Google.com, the authors argue that it is possible to address important graph mining tasks by analyzing the ego-nets of a social network and performing independent computations on them. The studies made by Everett and Borgatti [5] indicate that the local ego betweenness is highly correlated with the betweenness of the actor in the complete network. In [17] Wellman describes an ego network as a personal network. The author explains that the importance of local ties becomes apparent by redefining the composition of personal community networks in terms of the number of contacts (interactions) that egos have with the active members of the networks instead of the traditional procedure of counting the number of ties (relationships). In this work, we analyze the evolution of ego networks by using a bunch of social network analysis metrics.

We also discuss the growth pattern of our ego networks. In [8] the authors discussed how large graphs evolve over time. They stated a densification power law which is followed by these networks. In our work, we test the densification power law over the temporal ego networks of call graph stream and observe that it obeys the densification power law and follows the similar properties of large graphs. We also consider the properties of real world graphs depicted by [1],[2], [10], [16] such as diameter, path length etc.

As we observe the call ego networks satisfy the densification power law and the number of edges grow superlinearly to the number of nodes, the evolving graphs can get humongous in no time. There are a few sampling strategies discussed in [14] for sampling real time streaming graphs, but none of them preserve the tie strengths between nodes in the network. Nevertheless, there are no sampling techniques designed for ego networks to preserve the tie strengths, while maintaining the active and most recent nodes. Now the obvious question is, how do we capture the ego network of an evolving multi-graph stream over time with least possible edges, while preserving the structure, properties and efficiency of an ego network? For which, we proposed a streaming ego network sampling method using a forgetting factor [13]. The proposed method is suitable for dynamically evolving multi-graphs. We use this method over a real world temporal stream of edges/calls to generate a sample stream in real time. Our results show that the proposed method preserves tie strengths in the networks. We also show that our method decreases redundancy in the network while preserving the importance of ego. We measure the importance and efficiency of network using some socio-metrics. We evaluate our method by comparing the samples generated by varying parametric values, with the original ego network. The proposed method can also be implemented over a socio-centric network.

The following paper is organized as follows: In section 2 we discuss some related works. In section 3, we described our call network data and the metrics we used in our experiments in section 4. We proved the densification power law for our evolving ego networks in section 5. In section 6, we analyzed the properties and structure of evolving call ego network. Further in section 7, we proposed a sampling method for ego network multi graph streams with forgetting factor. Section 8 and 9, we evaluated the above method by comparing the samples with the original network.

2 Related Work

The concept of ego networks was discussed by L.C.Freeman in [6], where he described an ego network as a social network, built around a particular social unit called ego. In [17] Wellman discusses the importance of local ties in personal networks. In [3] Burt studied the affects, gaps and relationships between the neighborhood of a node, referring them as structural holes. He also introduced metrics to evaluate an efficient-effective network which strives to optimize structural holes in order to maximize information benefits.

Most of the research works in this field are carried out by analyzing the structure and growth pattern of evolving socio centric networks and evolutionary nature of socio centric graphs [1], [2], [10], [16], [11], [8]. Nevertheless, there are few works which studied the structure of ego networks [3], [7], [6], [13], [15]. To the best of our knowledge, this is the first work about analyzing the evolution of ego networks. We would analyze the evolution of ego network for 31 days using an extensive list of graph level and node level metrics.

[9] proposed an ego-centric network sampling approach for viral marketing applications. The authors employed a variation of forest fire algorithm for sampling ego network. They compared the degree and clustering coefficient distributions of sampled ego networks with the original ego network. In this work, we discuss an edge based sampling method with forgetting factor over an evolving ego network stream of temporal edges.

3 Description of Call Network Data

Telecommunications' call graphs are one of the massive streams of calls generated in real-time. We made use of such anonymised temporal call stream of 31 days available from a service provider. The network data stream is generated at a speed of 10 to 280 calls per second around mid-night and mid-day. On an average we have 12.4 million calls made by 4 million subscribers per day. Streaming approach is highly feasible for this kind of rapidly evolving data.

From the above massive stream of calls for 31 days, we built the ego networks by selecting egos with five different properties. The first ego network $egonet_1$ is built by selecting an ego with a degree equal to average degree of network. The $egonet_2$ with an ego of highest in-degree centrality of graph and is also the node with highest eigen vector centrality of graph. The $egonet_3$ and $egonet_4$

with highest betweenness centrality and lowest out-degree centrality respectively for enhancing the diversity of ego networks. We built these ego networks by accumulating all the adjacent edges of ego and their adjacent edges i.e a network of radius 2. We generated the ego network streams from a call network stream of 400 million calls made by 12 million subscribers on an aggregated scale. In order to avoid duplicated number of edges as it is a multi-graph, we maintained unique edges between any pair of nodes in the network and map them onto a weighted graph.

4 Metrics for Evaluating Ego Networks

In this section we discussed an extensive list of graph metrics which we would use in the later sections to analyze the evolution of structural, topological and behavioral properties and densification of call ego networks and to evaluate the proposed sampling method of forgetting factor. We exploit these properties at graph level and node level.

4.1 Graph Level Metrics

We studied the properties of ego network graphs using average degree, average weighted degree, density, diameter and average path length.

Additionally for evaluating evolving samples using our proposed method, we compared the degree distributions of the samples at the end of 31 days with the original network using kolmogorov-Smirnov test. We use the D-statistics from the test and also p-values to evaluate our null hypothesis (H_0) that our sampled ego networks follow the same distribution as the original ego network. The degree distributions of the networks is obtained by counting the frequency of each degree d in the network. The frequency of each degree d is given by the number of nodes with degree d in the network snapshots at the end of 31 days.

We compared the effective size and efficiency of samples with that of ego network using ego metrics introduced by Burt in [3]. Effective size of the ego network (ES) is the number of alters that an ego has, minus the average number of ties that each alter has to other alters. In the simplest form, for an undirected ego network of radius 1, the effective size can be given with the eq 1. Efficiency (EF) of an ego network is the proportion of ego's ties to its neighborhood that are "non-redundant." Efficiency is the normalized form of effective network size(eq 2). Therefore, it is a good measure for comparing ego networks of different sizes.

$$ES = n_a - \frac{\sum_{a=1}^{n_a}(d_a - 1)}{n_a} \tag{1}$$

$$EF = \frac{ES}{n_a} \tag{2}$$

where n_a is the number of alters in the ego network and d_a is the degree of an alter a.

4.2 Node Level Metrics

The node level centrality metrics discussed in the later sections are Degree, Weighted Degree, Closeness (CC), and Eigen Vector Centralities (EVC). We also explored the Eccentricity and Clustering Coefficient of the ego.

5 Densification Law over Evolving Call Ego-Networks

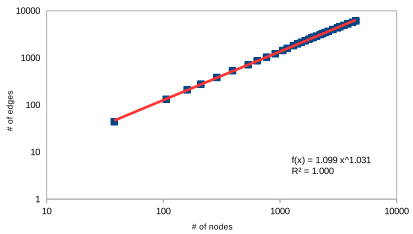
In [8] the authors studied the temporal evolution of, number of nodes vs number of edges. Besides, the authors employed the measures of average out degree to ascertain the densification law proposed by them. They validated that, most of these graphs densify over time, with the number of edges growing super-linearly to the number of nodes and their average degree increases. They investigated the above properties in an evolving citation graph, autonomous systems graph and affiliation graph. The authors stated that as the graphs evolve over time, they follow the relation given by the equation 3

$$e(t) \propto n(t)^a \tag{3}$$

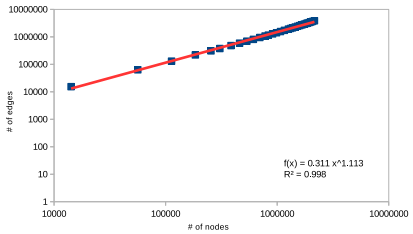
where e(t) and n(t) denote the number of edges and nodes of the graph at time t, and a is an exponent that generally lies strictly between 1 and 2. The authors refer to such a relation as a densification power law, or growth power law where the number of edges grow super-linearly to the number of nodes. The authors also show that the average degree of these graphs gradually increases. With this justification the authors prove that the graphs densify over time. In this section we investigate the densification power law (DPL) over the temporal stream of call ego network by depicting a densification power law plot (DPL plot) for the number of nodes n(t) and the number of edges e(t) at each timestamp t. In our experiments, we used a time stamp of one day.

We used the four temporal evolving ego networks from the call/edge stream as described in section 3. We grabbed the snapshots of ego networks at the end of each day and calculated the number of nodes and edges. Figure 1 shows the DPL plots for the call ego networks. As the slope of the line in a log-log plot gives the exponent in a power law relation, in the figure discussed above, we derived the lines obeying power relation with the best fits of 0.99 and 1.0 with their respective points. Therefore, the slope of these lines gives the densification exponents as a = 1.03, 1.1, 1.08, and 1.05 (in figure 1a, b, c and d respectively) which shows a super-linear growth of edges over nodes. Hence, we deduce that the ego networks of a call network also follow the densification power law as many other socio centric networks, with the number of edges growing super linearly to the number of nodes with their respective exponents a.

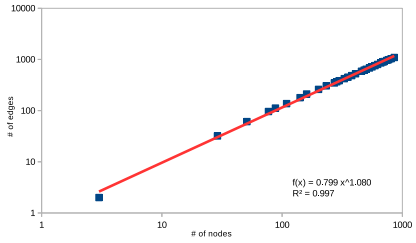
We consider the average degree of ego networks per time stamp, which is plotted in fig. 2. We see that the average degree of graphs for figure 2b and d (ego network of highest in-degree centrality node and highest eigen vector centrality of graph) is gradually increasing. Average degree of graphs in figures 2a and c are is slightly increasing with the evolution. From the above experiments with



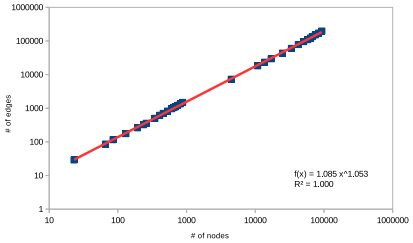
(a) $egonet_1$ evolution for 31 days, slope a=1.03



(b) $egonet_2$ evolution for 31 days, slope a=1.11

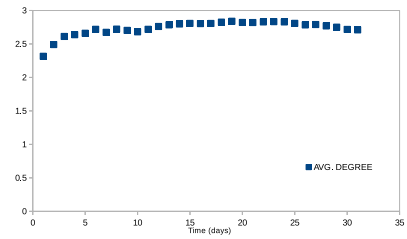


(c) $egonet_3$ evolution for 31 days, slope a=1.08

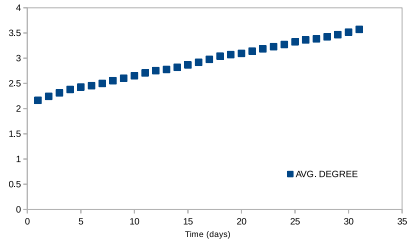


(d) $egonet_4$ evolution for 31 days, slope a=1.05

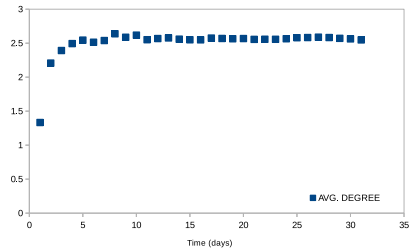
Fig. 1: DPL plot for temporal call ego networks



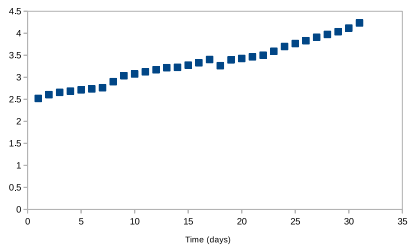
(a) $egonet_1$ evolution



(b) $egonet_2$ evoltuion



(c) $egonet_3$ evolution



(d) $egonet_4$ evolution

Fig. 2: Average degree evolution in temporal call ego networks for 31 days

the densification power law and the average degree, we see that the graphs are densifying. Hence we require sampling techniques in real-time to analyze such enormous evolving data.

6 Evolution Analysis of a Temporal Ego-Network

In this section, we have analyzed the evolution of an ego network ($egonet_1$) over a period of one month using the metrics mentioned in section 4. As described in earlier section, we have constituted the adjacent nodes of an ego and their adjacent nodes in the ego network as the stream progresses for a month. Then we took snapshots of the ego network per equal intervals of time stamp i.e one day in our case. To investigate the evolving structure and properties of a call ego network, we undertook a piecemeal structural analysis of network by employing the following metrics per day i.e. average degree, average weighted degree, Density, Diameter and Average Path Length and derive some empirical observations. We also made use of a bunch of centrality metrics to study the importance of ego in the network and compare the position of ego during evolution, they are degree, weighted degree, closeness (CC), and eigen vector centralities (EVC). We also explored the eccentricity and clustering coefficient of the ego.

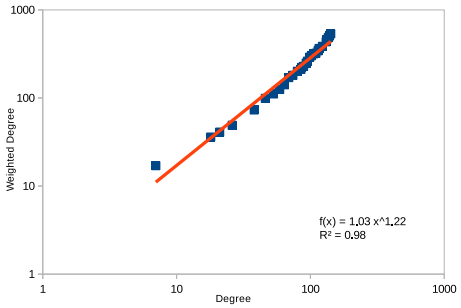


Fig. 3: Degree vs weighted degree of an ego network

Figure 3 plots the degree vs weighted degree of the ego in the ego network over a log log plot. The equation of the line that best fits our temporal data points is given in the figure. The slope of the line is given as 1.22 which shows a power relation between degree and weighted degree of a node. Therefore, we can say that the weighted degree of the node is growing superlinearly over its degree. The above analysis demonstrates a social behavior, that the people are more interested in maintaining their old relationships or friends than making new friends. However they also show interest in making new pals.

Figure 4 depicts the graph metrics and node metrics over an evolving ego network. When considering node metrics we see that the CC and eccentricity

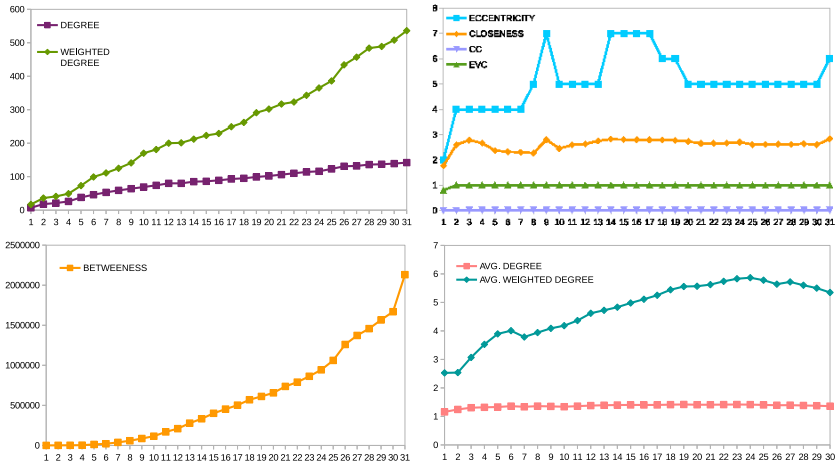
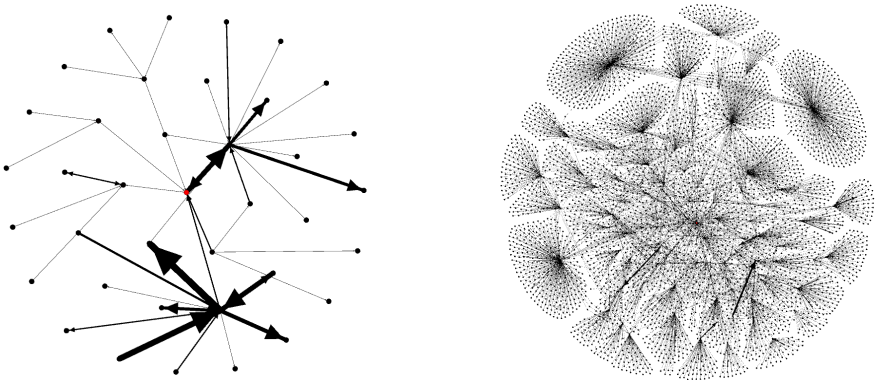


Fig. 4: Metrics over a temporal call ego network

of the ego increases with the evolution but EVC remains constant, as ego remains the important person in the network with highest betweenness centrality. Betweenness centrality of the ego also increases with the function of network size.

Figure 5 displays the call ego network evolution of a particular ego from day 1 to the final accumulated network on day 31. The ego is represented with a red dot in the center. We maintained the tie strengths between the nodes by mapping the multi graph to a weighted graph.



(a) (b)

Fig. 5: Metrics over ego networks with and without forgetting factor

7 Sampling Ego Network with Forgetting Factor (SEFF)

In this method, we sample edges from a stream of temporal network. We start by building the ego network of a specific ego and begin to scrape together all the adjacent ties to the ego and their adjacent ties. We do this by using a set for storing adjacent nodes. For every recurring edge, we increment the edge weight of the corresponding edge by maintaining a hash table. We impose a forgetting factor over edge weights, following successive time periods. In our experiments, we use a time period of 1 day. This means we apply the forgetting factor over the ego network as soon as the stream enters a new day, i.e we forget the weight of old edges $w(t-1)$, by some fixed percentage defined by the forgetting factor and sum it with the latest weight of edge w_t in time period t. The forgetting factor is given by two parameters, an attenuation factor α and a threshold θ. Where $0 < \alpha < 1$ and also $0 < \theta < 1$. After every time period t the tie strength between two nodes is given by a function $w(t)$ in the equation 4.

$$w(t) = w_t + (1 - \alpha) \times w(t-1) \tag{4}$$

After every successive time period, we decrease the edge weight by α and consequently remove the alter/alters adjacent to the corresponding edge, as the edge weight decreases than the threshold value θ. When α=1 we have a maximum forgetting i.e we forget the whole network except the network of current day. When $\alpha = 0$ we get the original network. If the removed edge corresponds to an alter adjacent to the ego, we remove the adjacent edge and the alter, and all the second level alters adjacent to the alter itself, if the above condition is satisfied. If we forget a second level edge, not having a direct connection to ego then we only forget the corresponding node. Following this strategy, we can have most active alters in the ego network at the end of each day.

8 Evaluation Methodology

In order to evaluate our method SEFF discussed in section 7, we applied it over a real world streaming call Graph G of 31 days by randomly choosing an ego e and generating a sample stream of depth $d = 2$ at any point of flow. This was done by generating six real time sample streams, where each sample stream S_i is generated by different combinations of $\alpha \in \{0.9, 0.8, 0.7, 0.5\}$ and $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ discussed in section 7. For investigating the above sample streams, we captured their snapshots of sample streams at the end of 31 days each. Each snapshot $S_i^{31} \subset G$. Beforehand, we took a snapshot G_e^{31} of original ego network stream G_e of e (where $d(G_e) = 2$) at the end of 31 days from the socio-centric call graph G. Each sample graph $S_i^{31} \subset G_e^{31}$. We then compared the conclusive sample snapshots S_i^{31} where $1 \leq i \leq 6$, with the original ego network snapshot G_e^{31} by employing metrics discussed in section 4. We use Kalmogorov-Smirnoff test to compare the degree distributions of the original network with that of samples. Conclusively, we derive some conclusions about the properties preserved by the sample networks.

9 Experimental Evaluation

The call networks are the special application scenario for employing our method as these networks are multi-graphs with more than one edge between two users, representing the strength of their relationship unlike a social network based on friendship and, follower and followee relations, where there is a single binomial relation between two nodes. However, the proposed method can be applied to networks with binomial relationships as it forgets edges and eventually forgets nodes. SEFF method is also appropriate for sampling weighted networks.

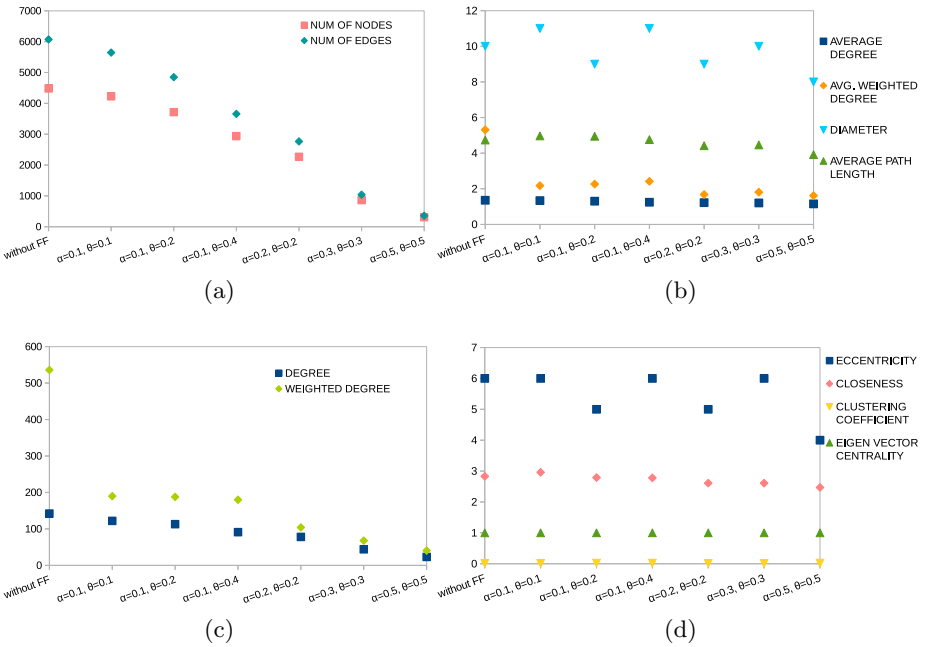


Fig. 6: Metrics over ego networks with and without forgetting factor

We selected an arbitrary user "ego" from the real world call/edge stream described in section 3 and start building the ego network of ego with a two step neighborhood, i.e. by acquiring the neighbors of ego and the neighbor of neighbors of ego. We take a snapshot of the ego network at the end of 31 days stream. Using the same ego we start constructing the sample ego networks (using SEFF) gradually as the stream flows for 31 days. For which, we have used six different combinations of α and θ corresponding to six different samples depicted in figure 6. The figure also plots the values of computed metrics discussed in section 4 over the conclusive sampled ego networks and the original ego network.

Figure 6(a) shows the number of nodes and the number of edges in the above described ego networks. We observe that the number of nodes gradually decrease with the increasing forgetting factor. For an attenuation value of 0.5 and threshold value of 0.5 we forget 50% of the edges per day, between two adjacent nodes. This shows we always have the most active nodes with the increased forgetting factor. We also observed that, the number of edges decrease in greater proportion than the number of nodes, Almost reaching equal for the highest forgetting factor in the illustration. This exhibits that the proposed SEFF method decreases redundant edges.

We also compare the degree distributions of the original ego network with the samples generated by using SEFF method at the end of 31 days. We applied Kolmogorov–Smirnov test to compare the degree distributions of the samples with the original network. The D-statistics and P-values of tests are given in the table 1. The p-values are computed using exact method. The significance level used for the comparisons is 5%, ie α=0.5. The results show that all the sampled distributions follow the distribution of original graph. We also observe that the value of θ has a greater impact on the similarity of distributions, than α in the SEFF method. We can see the pictorial representation of the degree distributions of the original graph and sample graphs in fig 7

Table 1: Comparison of degree distributions using KS-Test

Samples	α=0.1,θ=0.1	α=0.1,θ=0.2	α=0.1,θ=0.4	α=0.2,θ=0.2	α=0.3,θ=0.3	α=0.5,θ=0.5
D-stat	0.146	0.138	0.173	0.146	0.191	0.096
p-value	0.114	0.124	0.065	0.182	0.105	0.724

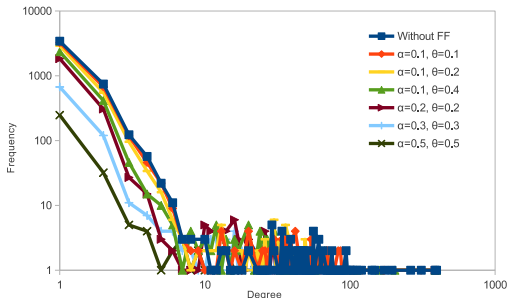


Fig. 7: Degree distributions of ego networks at the end of 31 days with and without forgetting factor

Figure 6(b) depicts metrics over the ego networks. The diameter of the graphs varies with the inclusion and removal of the connecting nodes from the ego

network. It depends on the network of ego selected. Average degree and the average path length decreases with the increasing forgetting, this shows that the networks shrink with increased forgetting. The SEFF method has a noticeable effect over the weighted degree of graphs.

The degree and weighted degree of the ego are plotted in figure 6(c). Both the values decreased with the increased forgetting, while the drop in weighted degree is higher, this suggests that when we increased forgetting we decreased the tie strengths but relatively maintained the ties. In Figure 6(d) we see that the eccentricity has a similar effect of diameter in the ego network graphs. This corresponds to the conceptual relation between diameter and eccentricity. Closeness of the ego with alters also decreased gradually with the increased forgetting factor. The clustering coefficient of ego is too low to compare. The eigen vector centrality portrays the important node in the network. SEFF preserves the importance of ego along side forgetting.

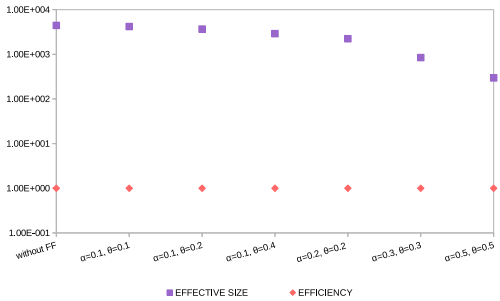


Fig. 8: Efficiency and effective size of ego networks

Figure 8 illustrates the effective size and efficiency of the ego networks. There is negligible difference in the effective size of samples. Efficiency of the network indicates the impact of ego in the network. In the given figure we can observe that the efficiency of the network is maintained through out the samples using SEFF. The measure of effective size of the network is not normalized with the size of network, therefore it decreases with the average number of ties that each alter has to other alters.

10 Conclusions

In this work, we analyzed the evolution of ego network for a period of one month. We exploited the structural properties of network and related them with the natural behavior of users. We also proved the densification law over the ego networks of call graphs for a period of one month and found that the graphs are densifying along time.

As we observed the properties of evolving ego network, we proposed a sampling method with forgetting factor for streaming multi-graph networks which preserves the density of graph and retains the tie strengths between nodes. We evaluated our method by exploiting the ground truth of original graph vs samples generated by varying parameter values. Based on the empirical experiments we prove that our method maintains the importance and efficiency of the network and decreases the redundancy while preserving most active and recent nodes from the network.

Acknowledgments. This work was partly supported by the European Commission through MAESTRA (ICT-2013-612944) and the Project TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/ NORTE-01-0145-FEDER-000020 is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement. Shazia Tabassum is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project (POCI-01-0145-FEDER-006961), and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The authors also thank WeDo Business for providing the data.

References

1. Albert, R., Jeong, H., Barabási, A.L.: Internet: Diameter of the world-wide web. Nature 401(6749), 130–131 (1999)
2. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer networks 33(1), 309–320 (2000)
3. Burt, R.S.: Structural holes: The social structure of competition. Harvard university press (2009)
4. Epasto, A., Lattanzi, S., Mirrokni, V., Sebe, I.O., Taei, A., Verma, S.: Ego-net community mining applied to friend suggestion. Proceedings of the VLDB Endowment 9(4), 324–335 (2015)
5. Everett, M., Borgatti, S.P.: Ego network betweenness. Social networks 27(1), 31–38 (2005)
6. Freeman, L.C.: Centered graphs and the structure of ego networks. Mathematical Social Sciences 3(3), 291–304 (1982)
7. Hanneman, R.A., Riddle, M.: Introduction to social network methods (2005)
8. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 177–187. ACM (2005)
9. Ma, H.H., Gustafson, S., Moitra, A., Bracewell, D.: Ego-centric network sampling in viral marketing applications. In: Mining and Analyzing Social Networks, pp. 35–51. Springer (2010)
10. Milgram, S.: The small world problem. Psychology today 2(1), 60–67 (1967)

11. Newman, M.E.: The structure and function of complex networks. SIAM review 45(2), 167–256 (2003)
12. Sarmento, R., Oliveira, M., Cordeiro, M., Tabassum, S., Gama, J.: Social network analysis of streaming call graphs. In: Big Data Analysis: New Algorithms for a New Society, vol. 16, pp. 239–261. Springer (2015)
13. Tabassum, S., Gama, J.: Sampling ego-networks with forgetting factor. In: IEEE Workshop on High Velocity Mobile Data Mining. p. In Press (2016)
14. Tabassum, S., Gama, J.: Sampling massive streaming call graphs. In: ACM Symposium on Advanced Computing. pp. 923–928 (2016)
15. Tabassum, S., Gama, J.: Social network analysis of mobile streaming networks. In: IEEE Conference on Mobile Data Mining, PhD Forum. p. In Press (2016)
16. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature 393(6684), 440–442 (1998)
17. Wellman, B.: Are personal communities local? a dumptarian reconsideration. Social networks 18(4), 347–354 (1996)